Safety Cases for Advanced Control Software: Final Report

Robert Alexander, Martin Hall-May, Tim Kelly and John McDermid
University of York
18th June 2007

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
24-09-2007	Final Report	20 September 2006 - 20-Sep-07

4. TITLE AND SUBTITLE

Safety Cases for Advanced Control Software

5a. CONTRACT NUMBER
FA8655-06-1-3041

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Professor John A McDermid

5d. PROJECT NUMBER

5d. TASK NUMBER

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of York
Heslington
York YO10 5DD
United Kingdom

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
PSC 821 BOX 14
FPO AE 09421-0014

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Grant 06-3041

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report results from a contract tasking University of York as follows: The project will undertake three activies:

1.Review current rules and regulations for clearing flight control software to establish a 'baseline' for the other two activies;

2. Assess the state-of-the-art in safety cases for adaptive systems and software, including neural networks and agents;

3. Outline a generic approach to developing safety cases for adaptive avionics and software.

Each activity would produce a stand-alone report for delivery to NASA.

15. SUBJECT TERMS
EOARD, software engineering, Reliability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18, NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			PAUL LOSIEWICZ, Ph. D.
UNCLAS	UNCLAS	UNCLAS	UL	15	19b. TELEPHONE NUMBER *(Include area code)* +44 20 7514 4474

1 Introduction

This project is concerned with the issues of assessing and flight clearing modern control applications – especially the software in such systems. Traditionally, software is cleared via a route which relies heavily on process evidence, and effectively requires that the behaviour of the software is pre-determined (if not deterministic). Modern control systems, especially for applications such as Unmanned Air Vehicles (UAVs), may use adaptive control algorithms, e.g. based on neural networks (NNs), which cannot realistically be cleared by the traditional approaches. The main focus of this work is on safety case approaches to dealing with such technology.

In practice, the notion of "adaptive" is rather fuzzy. A conventional control system which measures, say, air pressure and alters control accordingly is adapting to the environment – but we would not view this as being a truly "adaptive" system as we can say before operation what the behaviour will be. In general, a system is adaptive if its behaviour cannot be predicted from knowledge of the initial software design, the current (and recent) inputs to the system and knowledge of the hardware including failure status (e.g. unavailability of a sensor). An example might be a system based on NNs which learnt during operation and thus was able to compensate for damage to the wings which changes the aerodynamic properties of a UAV.

The work to date suggests that it is not practical to determine a "hard" boundary between adaptive and non-adaptive systems, but we can partly characterise the distinction, from the safety case perspective, by considering a simple example. The "engine" of an NN is deterministic – it is essentially an interpreter for a set of (learnt) rules. The engine could be developed to the requirements of a relevant standard, e.g. DO 178B [1], and a safety case presented that the NN met the requirements of the relevant level of DO178B (presumably level A for a flight control system). However, without evidence and arguments about the rules learnt by the NN we would not know what the control system would do. Thus even if the software part of the safety case met the requirements of DO178B, it would not be a compelling argument. We would have verification of the NN implementation, but no validation of the software behaviour. Thus a non-obvious, but it is hoped helpful, definition of adaptive systems would be those where conventional approaches to safety arguments could be misleading.

This report is the final report of this phase of the programme, however there is another short phase which is intended to complete the work. This report therefore summarises the work done to date and outlines the likely outcome of the work, in terms of possible safety case argument approaches, or patterns, for adaptive systems. As the work is not yet finalised, the discussion of the safety argument patterns is set out in general terms; full details, e.g. captured as GSN (Goal Structuring Notation) patterns, will be presented in the final report from the next phase of the work.

The core of this report is set out in section 2. It first summarises the work set out in the previous report, discusses relevant standards, and then considers approaches to safety arguments for adaptive control systems and software.

2 Progress

The Statement of Work (SoW) identified three tasks for the project. The first report [2] addressed items 1 and 2 on the SoW; this report addresses item 3

on the SoW, with the caveat set out above that the work will be completed in the associated project (effectively phase 2 of the work).

2.1 Literature Survey

The literature survey report addressed the first two items in the SoW:

1. Review current rules and regulations for clearing flight control software to establish a 'baseline' for the other two activies;
2. Assess the state-of-the-art in safety cases for adaptive systems and software, including neural networks and agents;

The report first addressed the problems of certifying adapative systems, considered some of the requirements of current safety standards, and then summarised current work on certifying adaptive technologies, such as NNs and Bayesian networks, in more detail. It also identified some work on clearance of novel control systems, especially flight control systems (FCS). As anticipated, the literature survey found a very limited set of results on the safety of adaptive systems and no substantive work on flight certification of agent-based technologies. The most mature work is in the area of NNs, including but not limited to work undertaken in York.

The report identified several possible approaches to assessing the safety of adaptive systems, including verification and validation techniques such as the use of formal methods. However there is relatively little established work on formal verification of adaptive systems, and there is a difficulty in establishing high fidelity formal models with a direct correspondence to the implementation of adaptive systems technologies and the environment in which they operate. Thus this does not seem like an attractive strategy. The survey concluded that it would be better to focus on product arguments (see below) including consideration of architectural mechanisms, e.g. safety monitors.

The literature survey report focused mainly on the requirements of the UK Defence safety standards not least because they offer the flexibility to produce “non-standard” arguments as part of a safety case. However it is important that the approaches to arguing safety are set in the context of standards more likely to be applied in the USA; this is done in the next sub-section before considering the approach to arguing safety.

2.2 Requirements of Standards

A discussion of the requirements of standards will be presented in the final report from the next phase. It is planned to address:

- DO178B (in more detail than in the literature survey)
- DO178C
- MilStd882

A comparison between MilStd 882 and DS 00-56 will also be presented, to link the discussion in the literature survey to current US Military Standards. There will also be a brief discussion of other standards, e.g. IEC 61508, to identify if there is relevant material in any of these standards.

2.3 Approaches to Arguing Safety

This sub-section is a preliminary response to the third item in the SoW:

3. Outline a generic approach to developing safety cases for adaptive avionics and software.

The strategy taken to address this item of the SoW is to identify principles and assumptions which can underlie and guide a safety case approach, to set out some top level arguments, then to identify some "alternatives" for lower level arguments. It is intended that, in the final report from the next phase, the principles and outline arguments set out here will be "fleshed out" to form safety case patterns. Note that this is a step towards a safety case "pattern catalogue" for adaptive systems; the set of patterns do not "join up" to form a complete argument but they provide building blocks from which arguments might be constructed.

2.3.1 Key Principles and Assumptions

As indicated above, process arguments are insufficient for adaptive systems. The strategy we advocate is to focus on the product and its properties, and this leads to the first principle:

> *Principle 1*: Safety arguments for adaptive systems must be primarily based on the product arguments and evidence, but supported by process arguments.

Even with a product focus, process arguments will still be necessary – perhaps addressing implementation verification, and certainly for confidence in the sufficiency of hazard identification. However, these arguments are (relatively) standard, so we will not amplify on them here.

When considering the product arguments, the main focus is on control over hazards – or perhaps better control of behaviour, including control over potentially unsafe behaviour. To articulate the next principle, it is helpful to consider (UAV flight) behaviour on a simple Venn diagram:

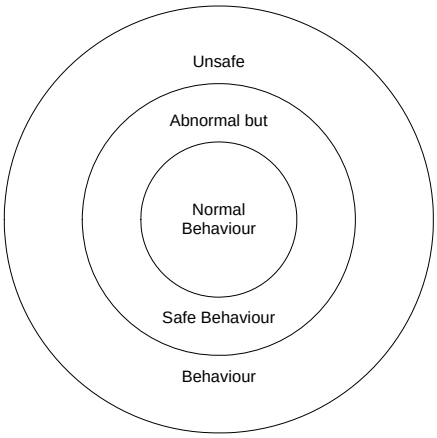

Figure 1: Simple Model of System Behaviour

For the purposes of this report, we make two initial assumptions:

> *Assumption 1*: Safe, normal behaviour could be obtained by conventional means, e.g. defining and implementing control laws.
>
> *Assumption 2*: The value of adaptive systems is primarily in that they extend the "abnormal but safe" region[1], and reduce the size of the unsafe behaviour region.

[1] There may be other benefits, e.g. better control or greater manoeuvrability; feedback from NASA on this principle would be appreciated.

These assumptions will prove to be important as they help to identify a number of safety case argument strategies. The first assumption can be used to help in articulating validation arguments. However it should be noted that this does not mean we are assuming hybrid implementations, e.g. using Matlab/Simulink for "normal" control modes, and NNs for abnormal situations – merely that a standard control model can be used as a basis for validating any design or implementation[2].

The second assumption will be central to any argument which needs to articulate benefits versus risks. Indeed it really identifies the "raison d'être" for adaptive technologies in flight control applications. An example of the benefit gained might be to be able to continue to fly a UAV following damage to a wing, by adapting the control algorithms to reflect the actual behaviour of the UAV, rather than its behaviour "as designed".

This assumption also gives rise to a further principle:

> *Principle 2*: The core product-based argument is that benefit accrues from the use of adaptive systems, without any increase in risk.

This principle could be formulated as a GSN pattern; as explained above, the definition of such patterns is deferred until phase 2.

Below this level there are a number of alternative ways of articulating arguments. The rest of this section identifies a number of key choices in approaches to argument construction – some of which are dependent on system design, and some of which are not. In practice the choice of argument strategies should be made in parallel with the system design. There is no implication that the order of concerns here reflects a decomposition of the safety argument – as stated above, the arguments should be viewed as the outline of a pattern catalogue.

There is a final assumption, relating to the use of adaptive technologies:

> *Assumption 3*: Classical safety analysis methods, including stochastic methods, will not be applicable to all parts of the operational envelope of the adaptive system.

This characterises the safety arguments that need to be constructed; at minimum it says that some of the arguments will need to be "non-standard" but it may be that classical approaches, e.g. using fault trees to demonstrate adequate risk control, are usable in parts of the argument.

2.3.2 Risk Control and Acceptance

There are several ways of arguing about risk control and presenting a framework for overall acceptance, but there seem to be two key alternatives:

> *RC1*: Adaptive system is at least as safe as a conventional system;
>
> *RC2*: The risk is reduced as low as reasonably practicable (ALARP).

Both of these arguments depend on the above Venn diagram showing the space of possible behaviours. We can amplify each argument as follows:

> *RC1*: Adaptive system is at least as safe as a conventional system:
>
> > *RC1G1*: Adaptive system behaviour is the same as a conventional system within the normal behaviour region;

[2] There is an implicit assumption here that testing, co-simulation, or similar will be sufficient to validate the adaptive control system. This may not be adequate and attention will be paid to this issue in finalising this report.

RC1G2: Adaptive system behaviour is the same as a conventional system within the conventional system's abnormal but safe behaviour region;

RC1G3: Adaptive system behaviour is safe for some of the region which is unsafe for the conventional system;

RC1A1: Adaptive system cannot impose behaviour intended to extend the abnormal but safe region so as to increase the risk in the normal or conventional abnormal but safe behaviour region.

RC2: The risk is reduced as low as reasonably practicable (ALARP):

RC2G1: Adaptive system provides a benefit of providing safe behaviour in the region which is unsafe for the conventional system;

RC2G2: The benefit of extending the safe abnormal behaviour region cannot be obtained without use of adaptive techniques;

RC2G3: The costs of further reducing the risks of using adapative systems exceed the benefits gained.

RC2G4: It has been possible to reduce the risks to a level that is tolerable.

In *RC1G1* and *RC1G2* we used the term "the same"; this should be read as meaning "the same within bounds that have a material impact on safe flight" or similar. In other words the behaviours do not have to be identical, merely close enough that they have no significant impact on flying qualities. It may also be necessary to assess what "the same" means in comparing piloted aircraft and UAVs, e.g. more manoeuvres may be deemed safe for a UAV, and there may be considerable difficulty in characterising human behaviour. We can think of *RC1G3* as being a "potentially safer than" goal.

For *RC2G3* an implicit assumption has been made that we are not in the "grossly disproportionate" region of the ALARP triangle. This assumption could be relaxed, but it is preserved for the rest of this document, as the form of ALARP arguments are well-understood, and little will be gained from adding complexity to this aspect of the argument (note: it does not change the structure of the argument, just parameters for the degree of disproportionality). The goal *RC2G4* is needed for a full ALARP argument.

In showing that *RC1G1* and *RC1G2* are met we need to carry out validation – given assumption 1, this could be by "back to back" assessment of the conventional and adaptive systems (see below). The argument and evidence for *RC1G3* is also a validation issue and will be dealt with in the same section.

Arguments and evidence in support of *RC2G1* and *RC2G2* can be built on Assumption 2. Support for *RC2G3* will depend on the specifics of the system and software development methods chosen, including the learning processes if NNs are used – and will necessarily link out to the process arguments (being concerned with effectiveness and costs of processes).

Assumption *RC1A1* should perhaps be articulated as a goal[3]; it also implicitly underpins *RC2G3*; again this is essentially a validation issue, but

[3] This change will be considered in developing the argument pattern catalogue.

there are some specific product issues related to this assumption – see below.

2.3.3 Control System Validation

Given Assumption 1 above, and general knowledge about control system design, we can see that there are essentially two different approaches that could be taken to validation:

> *V1*: Adaptive control system validated by "back to back" checks against conventional control laws (where applicable);
>
> *V2*: Adaptive control system validated by showing that it has adequate control and handling characteristics[4].

Essentially *V1* is an "at least as good as" approach, whereas *V2* is dealing with showing the adequacy of the adaptive control system in its own right. As with the discussion of *RC1* and *RC2*, these can be structured against the Venn diagram:

> *V1*: Adaptive control system validated by "back to back" checks against conventional control laws (where applicable):
>
> > *V2G1*: Adaptive control system validated by "back to back" checks against conventional control laws under normal behaviour;
> >
> > *V2G2*: Adaptive control system validated by "back to back" checks against conventional control laws under abnormal but safe behaviour;
> >
> > *V2G3*: Adaptive control system validated by simulation against damage hypotheses[5] which are in the unsafe behaviour region for the conventional control laws but which can be handled by the adaptive control system.
>
> *V2*: Adaptive control system validated by showing that it has adequate control and handling characteristics*:*
>
> > *V2G1*: Adaptive control system evaluated, ideally as a by-product of the design process, to show it has conventional handling qualities under normal behaviour;
> >
> > *V2G2*: Adaptive control system evaluated, ideally as a by-product of the design process, to show it has acceptable[6] handling qualities under abnormal but safe behaviour for conventional control laws;
> >
> > *V2G3*: Adaptive control system evaluated, ideally as a by-product of the design process, to show it has acceptable[7] handling qualities in the unsafe behaviour region for the conventional control laws but which can be handled by the adaptive control system.

---

[4] See *H Petrovski, To Engineer is Human*, for an interesting historical treatment of the way in which control engineers learnt to formalise handling qualities.

[5] If NASA has any guidance on the definition or validation of such hypotheses, that would be beneficial in developing this part of the argument pattern.

[6] This might be related to, say, the DO178B hazard categories, such as the description of the impact on safe flight (and crew workload), e.g. for Major.

[7] Again this might be related to the DO178B hazard categories, but for Hazardous.

These elements of the argument largely support *RC1G1* and *RC1G2* and/or *RC2G1* and *RC2G2*. An argument in support of *RC1A1* could be produced by showing that it does not affect control in the normal, or abnormal but safe regions if damage tolerance mechanisms are activated – or that these aspects of the algorithms cannot be activated. Some analysis pertinent to the latter approach is presented below (note that our aim here is to articulate key arguments, not to build a complete "joined up" argument). However some observations about "fleshing out" the argument can usefully be made.

V2G1-3 says "ideally as a by-product of the design process". Some of our earlier work on NNs, e.g. [3], had the very desirable characteristic of deriving necessary data for a safety case as the NN learnt. Figure 2 shows a fragment of the GSN for this approach:

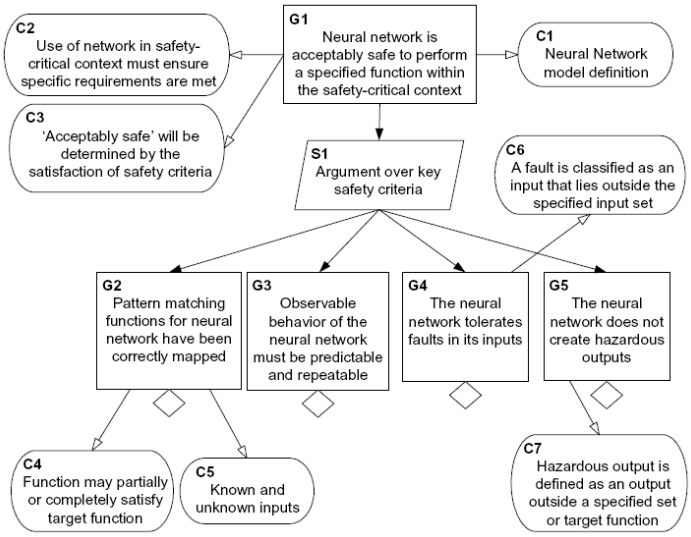

Figure 2: GSN for NN Development Process [3]

This both illustrates the nature of the top-level argument that might be used for an NN, but [3] also provides some backing to show that the phrase "ideally as a by-product of the design process" is not unrealistic for at least some classes of adaptive system.

Also, with regard to V2 (and again relating to the phrase "ideally as a by-product of the design process") some evidence can be given to show that this is not an unrealistic expectation (although detailed references can not be provided). One of the authors has seen a proposal for an aircraft fuel system employing NNs for fuel quantity indication (FQI). Here the proposed process was to provide training cases for the NN and to measure the error from the actual fuel quantity. The learning process was controlled statistically, to ensure that an upper bound could be placed on the error in the FQI value. It may be possible to carry out a similar training process for an NN-based control system so that it converged on acceptable handling qualities, within a well-controlled bound (see also the discussion of bounds and limits below).

We have used NNs as an example, as the technology is relatively mature, however there are other interesting approaches, including use of heuristic optimisation in control system design. There is a wide range of literature on the use of heuristic approaches to controller design. There are not well-known solutions to show that the results of these heuristics are dependable. However an interesting possibility is to place trust in the value function used

to evaluate designs produced by the heuristic search – these value functions are usually deterministic, and represent explicitly "figures of merit" for the control system design. This is another example that shows that it is not unreasonable to use the phrase "ideally as a by-product of the design process" in the above goals. There is, however, an issue of when the control optimisation (or other form of adaptation) is done – specifically whether it is at design time or operation, see below for further discussion of this issue.

The outline argument patterns in sections 2.3.2 and 2.3.4 are quite strongly connected; the following outline patterns are less strongly connected, and should be viewed as suggesting the basis of a pattern catalogue, and there is no attempt to present a "joined up" argument.

2.3.4 Bounds, Limits and Monitoring

One general approach to achieving safety is to place bounds, or limits, on system actions so that the system as a whole, e.g. a UAV, stays within a safe operating envelope. This might mean, for example, employing a rate limiter on an actuator, or limiting the amount of time an actuator operates at maximum rate[8]. This seems potentially attractive for adaptive systems, by offering the ability to adapt within known safe bounds – with the weight of the safety argument on the bounding behaviour, which might be achieved by a separate monitor function.

The examples above suggest that limits are placed on individual actuators. However, for the class of applications being considered here, this might actually prevent the "extension" of the abnormal but safe region through the use of adaptive systems. This leads us to a further assumption (strictly two alternative assumptions):

> *Assumption 4a*: Limits on behaviour of an adaptive control system need to be determined in terms of aircraft dynamics, in order not to lose the benefits of employing adaptive control.
>
> *Assumption 4b*: Limits on behaviour of an adaptive control system can be applied hierarchically, with the highest authority given to those limiters operating in terms of aircraft dynamics, in order not to lose the benefits of employing adaptive control.

Under assumption 4b, for example, an actuator rate limit, or the maximum allowed differential in control surface position, might be over-ridden by an aircraft level limiter function. Note that the above implicitly assumes the ability to determine enough about the UAV trajectory to be able to determine how to apply limits; it is outside our technical competence to judge whether or not (to what extent) this is possible, and this is an issue on which we would appreciate guidance from NASA. For simplicity here we work on the basis of assumption 4a – which we can do without loss of generality[9].

There are two alternative design approaches to achieving bounding or limiting behaviour – that it is intrinsic in the design, i.e. "built in" to the control mechanisms, or separate, typically using a "monitor". Thus the choice of argument approach is driven by the system architecture and

---

[8] Such limits are used in some current commercial aircraft control systems.

[9] Also, this may be more practical, as designing a hierarchical limiter may be quite difficult as the "higher level" limiters may need to over-ride (countermand) the "lower level" limiters, see also the section below concerned with observability of events.

detailed algorithm design. As a consequence there are two alternative starting points for this aspect of any argument:

B1: Adaptive control system intrinsically stays within safe bounds;

B2: Adaptive control system is monitored and the outputs mediated to ensure that the overall control system stays within safe bounds.

The focus of the two approaches is rather different – the former focused on the adaptive controller itself, and the second focused on the monitor. However both rely on a common assumption (and perhaps should be modified to reflect the assumption):

Assumption 5: There are levels of system loss or UAV damage which cannot be contained by any control system.

This loss is referred to as "extreme loss of capability" and will always lead to a transition into the unsafe behaviour region. The argument strategies are:

B1: Adaptive control system intrinsically stays within safe bounds:

B1G1: Adaptive control system reduces (ideally eliminates) the deviation between UAV dynamics and intent and keeps behaviour within safe bounds, up to an extreme loss of capability.

B2: Adaptive control system is monitored and the outputs mediated to ensure that the overall control system stays within safe bounds:

B2G1: Monitor can identify deviation of aircraft dynamics from intent;

B2G2: Monitor can modify actuator commands from the (adaptive) control system to reduce (ideally eliminate) the deviation between UAV dynamics and intent and keep behaviour within safe bounds, up to an extreme loss of capability.

Arguably the early part of *B1G1* and *B2G2* are unnecessary, as they essentially state the purpose of a control system – they are included here as they relate to overall vehicle control, and cannot be limited to "local" control, e.g. governing actuator position and rate of movement.

The argument structure at *B2G2* suggests that the control system may be "classical". Indeed, with a controller-monitor structure it may well be appropriate to use a classical controller and to focus the use of adaptive control technologies on the monitor. However, this design strategy depends on the ability to detect or observe deviation from the normal control regime, see below. Further, as the monitor will have to be high authority this may mean that the certification argument becomes more complex; as with other aspects of the proposed argument pattern catalogue, there are some important issues of detail to consider when "fleshing out" the patterns. If this does prove difficult it may indicate that "strategy" *B1* is preferable.

The argument structure here is quite shallow – but that seems unavoidable. It may be that technology specific arguments can be produced that take this part of the argument pattern further; the NN argument at figure 2 is an example of what a technology specific argument might encompass.

2.3.5 On-line versus Off-line Adaptation

It seems that a significant variation in the "certification challenge" arises from the choice between off-line and on-line adaptation. In the off-line case, all adaptive capability is defined during the development process; in the on-

line case adaptation can occur whilst the system is operating. The on-line case is the more challenging, but also the more likely to be of value in dealing with extensive UAV damage.

An off-line design approach can deal with "defined" situations, e.g. sensor failures and assumed damage scenarios. An approach such as the use of a non-standard form of NNs set out in [3] can be used in this case. The capability of the system will depend on the representativeness of the training set to situations which are encountered in practice. (The behaviour is not limited to the training set, as the NNs will interpolate between these points.)

Intuitively the on-line approach should be more effective, as it can deal with situations which actually arise, not just those imagined in the off-line training activity (and those interpolated by the NN). Note, however, that this remains a hypothesis, but it is outside the scope of this study[10] to investigate this hypothesis in detail. This approach would seem to be harder to assure than the off-line approach, essentially because the behaviour of the controller system is not known (fully) before operation. However the approach adopted in [3] overcame this problem by assuring the bounds or limits, rather than the behaviour as such; thus this remains a potentially attractive approach, with some adaptive technologies.

Thus there are two alternative strategies for this aspect of an argument:

> *OL1*: On-line control system adaptation ensures safe operation maximising the size of the abnormal but safe operating region;
>
> *OL2*: Off-line control system design produces safe control system.

At this level of analysis *OL2* really "degenerates" to *V1* or *V2* plus the necessary verification activities. Goal OL1 is much more challenging:

> *OL1*: On-line control system adaptation ensures safe operation maximising the size of the abnormal but safe operating region:
>
> > *OL1G1*: see *V1* or *V2,* i.e. there is a need for validation;
> >
> > *OL1G2:* Adaptive behaviour does not compromise the safety of the behaviour in the normal and abnormal but safe regions;
> >
> > *OL1G3:* Adaptive behaviour maximises the damage/failure scenarios that can be accommodated in the abnormal but safe region;
> >
> > *OL1G4:* Control system implemented as specified, and evolution cannot compromise the integrity of the implementation of the "normal" control functionality.
>
> *OL2*: Off-line control system design produces safe control system:
>
> > *OL2G1*: see *V1* or *V2,* i.e. there is a need for validation;
> >
> > *OL2G2:* Control system implemented as specified.

In practice *OL2G2* would link into a "classical" verification activity, perhaps using a DO178B accomplishment summary, or similar.

The arguments in support of *OL1* will depend on the system architecture. For example, *OL1G4* may be easy to argue if the implementation is a hybrid "classical" control system with an adaptive monitor, and the two subsystems are effectively partitioned. Goal *OL1G2* will be easier to support in the design and in the safety case if it is possible to reliably detect transitions between

[10] And the skills of the authors.

the regions in the Venn diagram in figure 1. If this can't be done, then the adaptive rules will have (must have) authority throughout the flight envelope – but this might be possible to discharge by using the evidence in support of *V2G1* and *V2G2*.

It is likely that the best that can be done for *OL1G3* will be to use simulation and to argue that the failure and damage scenarios used to validate the approach are credible and challenging. This can perhaps be "measured" in terms of the distance towards the "extreme loss of capability" identified above. In some sense, the ideal design would keep the UAV flying until the boundary of the "extreme loss of capability", then "immediately" lose control. Thus a "figure of merit" would be a measure of how far the control system gets from the boundary of the abnormal but safe region (with conventional control) towards the boundary of the "extreme loss of capability" – recognising that this is not a linear progression, and there is unlikely to be a simple measure of this "figure of merit".

Note that the discussion here overlaps somewhat with the discussion of *V1* and *V2* in section 2.3.3. When fleshing out the patterns, it may necessary to modify the boundaries of these two argument pattern fragments to address the issues most effectively; assuming both arguments form part of the final set of patterns it is likely that they will be "collaborators".

2.3.6 Observability or Detectability of Events and States

The observability, or detectability, of certain events and states are important to some of the argument patterns outlined above. In particular they will be important where the controls are moded, perhaps particularly if the mode change involves moving from classical to adaptive control (doing this at the wrong time might violate some of the safety goals, e.g. *OL1G2*.)

As we are concerned with flight controls that can preserve safe flight through failures and damage, it is clear that not all events of interest can be sensed directly, e.g. we will not, in general, be able to sense damage. However, as we are mainly concerned with overall control of flight (consider for example the notion of a monitor), and extending the abnormal but safe region, it is perhaps more important to think in terms of detecting transition between the regions. At best, this would be an inferred transition – as the "boundary" will be complex and will not be measurable by any single sensor. Thus the issue is perhaps more one of being able to infer transition between regions of flight which are significant from a control perspective.

There are two alternative starting points for this aspect of any argument:

> *D1*: Mode transitions which are significant for safe control can be detected or inferred from sensor data and the control system makes appropriate changes in control regime in response;
>
> *D2*: Mode transitions which are significant for safe control cannot be detected or inferred from sensor data and the control system provides appropriate control without explicit mode detection.

At first sight *D1* may prove preferable, as it means that the control system can detect "crossing" the boundaries of the regions in the Venn diagram in figure 1. In contrast, *D2* has the apparent advantage of robustness against loss of some classes of sensor data although a detailed analysis would need to be done to show that sensors needed to enable flight control weren't also needed for detection of boundary crossing. If that proves to be the case, then *D2* probably does not offer any real advantages.

The goals could be expanded as follows:

D1: Mode transitions which are significant for safe control can be detected or inferred from sensor data and the control system makes appropriate changes in control regime in response:

D1G1: Sensors used for flight control can be used to detect or infer the boundary of the safe control region, e.g. stall or excessive yaw;

D1G2: The control system enters appropriate control modes sufficiently quickly[11] to enable control action to be taken before the UAV enters the unsafe region;

D1G3: The sensing of boundary crossing is sufficiently accurate to prevent (reduce to an acceptable likelihood) false activation of abnormal control modes (see *RC1A1*).

D2: Mode transitions that are significant for safe control cannot be detected or inferred from sensor data and the control system provides appropriate control without explicit mode detection.

D2G1: The adaptive control system has sufficient capability (and authority) to operate in "closed loop" control against sensed flying qualities without the need to detect boundary crossings.

Strategy *D2*, as should be apparent from *D2G1*, most naturally aligns with *V2* in section 2.3.3.

2.3.7 Controllable versus Uncontrollable outputs

It may also be the case that the "controllability" of outputs is significant, where we use controllability to mean the ability to directly influence a property of interest, e.g. the position of a control surface, or the angle of attack of a UAV. However, at this stage it is unclear that this will have a major influence on the arguments (as it is equally an issue for classical control mechanisms) so the issue is not amplified, at this stage.

2.3.8 Observations

It is normal to try to construct design patterns or safety case patterns by assessing a range of successful designs or arguments, generalising and documenting commonly occurring reusable "fragments". Due to the absence of safety cases for adaptive control systems, and the relatively limited literature on the subject, it has not been possible to adopt this approach. Instead the construction of arguments has been approached "from first principles" and a set of possible patterns has been sketched out. Considerable judgment has been applied in constructing these outline patterns – it is thus highly improbable that all the proposed patterns will "stand the test of time", but it is hopefully clear that they do address at least some key issues.

These proposed patterns are obviously related although they do not all "join up". It is not realistic to expect an argument pattern catalogue to be complete in this sense, as many details will depend on the design approach chosen, e.g. if Bayesian networks are used, then low level arguments will be about defining weights in conditional probability tables, and the adaptation

[11] If on-line adaptation or even optimisation is used, then this is likely to be a particularly challenging goal to meet.

of the control system (see *OL1G2*) would relate to on-line modification of these tables. However, it is hoped that much of the top level of the safety case for an adaptive flight control system has been identified in the set of patterns and that they will provide a framework for such lower-level arguments.

The patterns will also serve a useful purpose in providing a framework within which some of the current difficulties of assuring adaptive system behaviour can be highlighted, discussed and understood. It is thus hoped that in constructing these patterns some insight will be gained into what are likely to be the most fruitful ways of assuring safety of adaptive FCS.

3 Conclusions

The aim in this report has been to sketch out possible approaches to arguing about the safety of adaptive control systems. The earlier analysis in the project indicated that a product-focused argument would be needed. The report has set out some outline arguments necessary to provide this product-focused argument. No attempt has been made to produce a "joined up" argument, as the details of such an argument will depend on the particular system design being considered. Instead the aim has been to produce an initial definition of the scope and contents of a catalogue of relevant safety argument patterns.

The next and final phase of the work is intended to turn the outlines above into patterns in GSN. When this is done, it is likely that the details of some of the sketched patterns will change and some omissions may also become apparent. Some enrichment of the patterns are also likely to be needed and this will include reworking some of the textual comments into "formal" parts of the pattern structure, e.g. that (the patterns supporting) *D2* and *V2* are collaborators.

It is also worth making a caveat on these proposed argument patterns. Experience with safety cases and elsewhere is that it is generally best to produce patterns by reviewing existing material. However, as there are, as yet, no existing safety cases for adaptive flight control systems it is not practical to follow this course. Thus it must be made clear that these proposed patterns are more tentative than would be the case if they had been derived from practical experience. However, use of some of the patterns implies some architectural constraints. Thus it is hoped that what is set out here will provide some useful input to designers of future adaptive control systems, and that it will help them think about approaches to certification, as well as providing some building blocks for safety case developers.

4 References

[1] DO178B, *Software Considerations in Airborne Systems and Equipment Certification*, EUROCAE/RTCA, 1992.

[2] R Alexander, M Hall-May, T P Kelly, J A McDermid, Clearance of Learning and Adaptive Systems for Safety-Critical Applications – a Literature Review, University of York, January 2007.

[3] Z Kurd, T P Kelly, J Austin, Developing Artificial Neural Networks for Safety Critical Systems, Neural Computing and Applications, 16: 11-19, 2007.